UC Berkeley

UC Berkeley Previously Published Works

Title

Characterization of changes in gene expression and biochemical pathways at low levels of benzene exposure.

Permalink

https://escholarship.org/uc/item/3jj827p5

Journal

PloS one, 9(5)

ISSN

1932-6203

Authors

Thomas, Reuben
Hubbard, Alan E
McHale, Cliona M
et al.

Publication Date

2014

DOI

10.1371/journal.pone.0091828

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

OPEN ACCESS Freely available online

PLOS ONE

Characterization of Changes in Gene Expression and Biochemical Pathways at Low Levels of Benzene Exposure

Reuben Thomas[1]*, Alan E. Hubbard[1], Cliona M. McHale[1], Luoping Zhang[1], Stephen M. Rappaport[1], Qing Lan[2], Nathaniel Rothman[2], Roel Vermeulen[4], Kathryn Z. Guyton[3], Jennifer Jinot[3], Babasaheb R. Sonawane[3], Martyn T. Smith[1]

1 Superfund Research Program, School of Public Health, University of California, Berkeley, California, United States of America, **2** Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **3** National Center for Environmental Assessment, Office of Research and Development, US EPA, Washington, DC, United States of America, **4** Institute of Risk assessment Sciences, Utrecht University, Utrecht, The Netherlands

Abstract

Benzene, a ubiquitous environmental pollutant, causes acute myeloid leukemia (AML). Recently, through transcriptome profiling of peripheral blood mononuclear cells (PBMC), we reported dose-dependent effects of benzene exposure on gene expression and biochemical pathways in 83 workers exposed across four airborne concentration ranges (from <1 ppm to > 10 ppm) compared with 42 subjects with non-workplace ambient exposure levels. Here, we further characterize these dose-dependent effects with continuous benzene exposure in all 125 study subjects. We estimated air benzene exposure levels in the 42 environmentally-exposed subjects from their unmetabolized urinary benzene levels. We used a novel non-parametric, data-adaptive model selection method to estimate the change with dose in the expression of each gene. We describe non-parametric approaches to model pathway responses and used these to estimate the dose responses of the AML pathway and 4 other pathways of interest. The response patterns of majority of genes as captured by mean estimates of the first and second principal components of the dose-response for the five pathways and the profiles of 6 AML pathway response-representative genes (identified by clustering) exhibited similar apparent supra-linear responses. Responses at or below 0.1 ppm benzene were observed for altered expression of AML pathway genes and *CYP2E1*. Together, these data show that benzene alters disease-relevant pathways and genes in a dose-dependent manner, with effects apparent at doses as low as 100 ppb in air. Studies with extensive exposure assessment of subjects exposed in the low-dose range between 10 ppb and 1 ppm are needed to confirm these findings.

Citation: Thomas R, Hubbard AE, McHale CM, Zhang L, Rappaport SM, et al. (2014) Characterization of Changes in Gene Expression and Biochemical Pathways at Low Levels of Benzene Exposure. PLoS ONE 9(5): e91828. doi:10.1371/journal.pone.0091828

Editor: Shyamal D. Peddada, National Institute of Environmental and Health Sciences, United States of America

Received June 11, 2013; **Accepted** February 14, 2014; **Published** May 1, 2014

Copyright: © 2014 Thomas et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Original data generated through funding from National Institutes of Health grants R01ES01896 and P42 ES004705 from the National Institute of Environmental Health Sciences with additional statistical analyses being funded in part by Environment Protection Agency contract number EP-11-001398. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: S.M.R. has received consulting and expert testimony fees from law firms representing plaintiffs' cases involving exposure to benzene and has received research support from the American Petroleum Institute and the American Chemistry Council. M.T.S. has received consulting and expert testimony fees from law firms representing both plaintiffs and defendants in cases involving exposure to benzene. The other authors declare they have no actual or potential competing financial interests. Alan Hubbard is an Associate Editor of this journal. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: reuben.thomas@berkeley.edu

Introduction

Benzene is a component of gasoline, and the starting ingredient in the production of plastics and polymers via styrene; of resins and adhesives via phenol; and, in the manufacture of nylon via cyclohexane. It is toxic to the bone marrow and is associated with various hematological cancers [1,2].

Multiple possible mechanisms of action are thought to be involved in benzene toxicity [3,4,5,6]. Benzene exposure has been shown to cause hematotoxicity [7], induce formation of protein adducts [8,9], and increase the risk of leukemia [10], in a dose-dependent manner. Linear or supra-linear dose-dependent effects on lymphocyte counts and colony formation from myeloid stem and progenitor cells and gene expression were reported at relatively low levels of occupational exposure (≤1 ppm to > 10 ppm) in exposed human populations [7,11,12]. Recently, through transcriptome profiling of peripheral blood mononuclear cells (PBMC), we reported dose-dependent effects of benzene on gene expression and biochemical pathways in 83 workers exposed to air benzene levels across four concentration ranges (from < 1 ppm to >10 ppm), compared with 42 subjects not occupationally exposed to benzene [13]. A 16-gene signature associated with all levels of benzene exposure exhibited an apparently supra-linear dose response. In addition, several immune response-related pathways and the pathway associated with AML were significantly modulated across several of the benzene dose ranges examined.

A deeper understanding of the dose-dependent, disease-relevant human biochemical responses resulting from benzene exposure, particularly at low doses, is important for next generation approaches to human health risk assessment. Therefore, the goal

of the current study was to further characterize the dose-dependency of low-dose effects of benzene on genes and biochemical pathways identified in our recent benzene-related microarray analyses of PBMC [13]. Specifically, continuous data for individual benzene exposure across all dose groups was used to generate dose-response curves on a continuous scale. We used predicted measures of benzene exposure in the group of subjects with only ambient exposure to benzene. These subjects were regarded as controls in the previous study but were, in fact, non-occupationally exposed to benzene at varying, relatively low environmental concentrations. Inclusion of data from these individuals allowed us to examine more closely the responses in the low dose (environmental) region of exposure. Using data from all 125 study subjects, we applied non-parametric approaches, based on the SuperLearner [14], to fit the responses of individual gene expression as a function of benzene exposure. The use of non-parametric approaches is particularly relevant here and in epidemiological studies in general because it is impossible to know the exact functional relationships among the variables such as gene expression, dose from exposure, age, gender and smoking status of the subject, cell counts etc. Non-parametric approaches make minimal assumptions about these functional relationships and let the observed data guide the choice of the best models using rigorous statistical criteria (*e.g.*, cross-validation [14]). The implication of making parametric assumptions is that if these assumptions are untrue (which is almost certainly the case), the results produced can be difficult to interpret. In the current study, we developed novel non-parametric approaches to model the responses in biochemical pathways of interest. We chose to model the responses in 5 pathways, including the AML pathway and two other pathways previously shown to be modified by benzene, and two pathways presumably unrelated to benzene exposure. We also employed the models to examine dose-response relationships in the expression of a set of candidate genes known to be associated with AML and with the metabolism of benzene.

The overall goals of this study were to estimate the benzene exposure-response patterns of relevant gene expression and biochemical pathways in a statistically rigorous, non-parametric manner. This approach allowed us to identify consistencies in the shapes of the resulting exposure-response curves and characterize responses particularly in the low-dose region of exposure. Since our original microarray data were generated from PBMCs which comprise various cell types [15], including T lymphocytes (CD4 and CD8 ~65%), B cells (~15%), natural Killer cells (~10%), and monocytes (~10%), we adjusted for changes in percentages of these subtypes after benzene exposure in our analyses.

Materials and Methods

A brief overview of the data and methods used is given in Figureô 1.

Data Sets

Ethics statement. This study complied with all applicable requirements of U.S. and Chinese regulations, including institutional review board approval at the National Cancer Institute, Bethesda, Maryland USA and the National Institute of Occupational Health and Poison Control, China CDC, Beijing. Participation was voluntary, and written informed consent was obtained.

Study population, hematotoxicity, and gene expression data. The overall molecular epidemiology studies investigating occupational exposure to benzene [7,16] and the gene expression data [13] upon which the analyses in the current study are based were previously described. The gene expression data were generated through transcriptome analysis by microarray of 125 subjects exposed to various levels of benzene. Among the 125 subjects, 42 were exposed to levels that were below the limit of detection of the benzene monitors (0.04 ppm) used; 29 were exposed to $<<1$ ppm (average <1 ppm and most individual measurements <1 ppm) benzene; 30 were exposed to <1 ppm (average <1 ppm); 11 were exposed to levels between 5 ppm and 10 ppm; and 13 were exposed to levels ≥ 10 ppm. For each of the exposed individuals in the study, benzene exposure was estimated in terms of the average air-benzene level (in units of parts-per-million). The exposure levels of the 42 subjects that were below the limit of detection were estimated using unmetabolized urinary benzene levels, as previously described [17]. Complete blood cell counts, including counts for CD4 and CD8 T lymphocytes, B lymphocytes, NK cells and monocytes, the major cell subsets of PBMCs, were available for all the individuals analyzed by microarray [7].

Biochemical Pathways

The biochemical pathways analyzed in this study were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database [18,19,20]. The data for the set of genes within each pathway and their associated interactions were downloaded using the KEGG application programming interface (http://www.kegg.jp/kegg/soap/). Five pathways were analyzed, including three (AML, B-cell receptor signaling and Toll-like receptor signaling) previously shown to be differentially modulated with benzene exposure [13] and two (Steroid hormone biosynthesis and Maturity onset of diabetes) presumably unrelated to benzene exposure were not differentially modulated.

Linear Mixed Effect Models

We conducted variance components analysis using a linear mixed model [21] to assess the proportion of total variation due to differences between subjects, hybridizations, labels, and chips, both before and after normalization [quantile normalization in the Affy package [22] in R [23]]. For each probe, we estimated the association between exposure level and expression level using a mixed-effects model with random intercepts that accounted for clustering by subject, hybridization, and label. The fixed effects in our model included gender (1 = male, 0 = female), current smoking status (1 = yes, 0 = no), age (in years, linear term), B cells, Natural Killer (NK) cells, monocytes, and CD4 and CD8 cells (as counts and included as linear terms). These were potential confounders of associations (denoted by the vector of random variables, W) between logarithm to the base 2 of gene expression (denoted by the random variable, Y) and benzene exposure in five dose ranges (denoted by the random variable, A). The model is thus given by,

$$\begin{aligned} Y^g_{ijklm} = {} & \beta^g_0 a_i + \beta^g_1 + \beta^g_2(sex_j) + \beta^g_3(smoke_j) + \beta^g_4(age_j) \\ & + \beta^g_5(\#B\,cells_j) + \beta^g_6(\#NK\,cells_j) + \beta^g_7(\#monocytes_j) \\ & + \beta^g_8(\#CD4\,cells_j) + \beta^g_9(\#CD8\,cells_j) + \mu^g_j(subject) \\ & + \mu^g_k(hybridization) + \mu^g_l(label) + \varepsilon^g_{ijklm} \end{aligned} \quad (1)$$

Y^g_{ijklm} denotes the $\log_2$ of the g^{th} gene expression, at the dose a_i, $a_i \in \{0,1,2,3,4\}$ obtained from the j^{th} subject after the k^{th} hybridization, l^{th} labeling step in the microarray sample preparation and the m^{th} chip. The β^g parameters denote the fixed effects associated with the respective covariates; the μ parameters denote

Outcome (Y): Gene expression in PBMCs[a]

Predictor (A): Air borne benzene levels

Confounders (W): Gender, Smoking Status, Age, Counts of PBMCs

Data: Gene expression in PBMCs of 125 subjects exposed to a range of benzene levels

Parameter: Benzene-dose dependent expected fold change in $\log_2$ expression of each gene relative to the expression in control or very lowly exposed subjects (<0.11ppm) margining out the effects of the other confounders

Analyses:

	Parametric	*Non-parametric*
Dose	5 binned dose ranges	Continuous
Point estimation	Linear mixed models	SuperLearner
Gene parameter	Fixed effect corresponding to dose	Ψ_i^g (see equation 2)
Gene inference	Linear mixed models	SuperLearner + Bootstrapping
Biochemical pathway inference	Linear mixed models + SEPEA[d]	a. SuperLearner + Bootstrapping + PCA[b]
		b. SuperLearner + Bootstrapping + HOPACH[c]

[a]PBMC: Peripheral Blood Mononuclear Cells; [b]PCA: Principal Component Analyses; [c]HOPACH: Hierarchical Ordered Partitioning and Collapsing Hybrid; [d]SEPEA: Structurally Enhanced Pathway Enrichment Analyses

Figure 1. Overview of methods and analyses.
doi:10.1371/journal.pone.0091828.g001

the random effects, and ε denotes the normally distributed error associated with the model. β_0^g, the fixed effect associated with benzene exposure, is the parameter of interest in the model. We fitted this mixed-effects model in R with the lmer function in the *lme4* package [24]. We also fit the mixed effects model without cell counts as potential confounders as given in Equation (2).

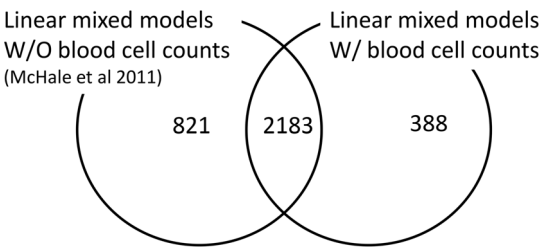

Figure 2. Overlapping sets of genes determined by two linear models. Two linear mixed models were used, a published model [13] (see Equation (2)) and a modified version including counts of different blood cell types as potential confounders of gene expression (see Equation (1)). Differential expression was determined based on altered fold changes in at least one of the four previously chosen dose ranges of benzene exposure, with an FDR-adjusted p-value<0.05.
doi:10.1371/journal.pone.0091828.g002

$$Y_{ijklm}^g = \beta_0^g a_i + \beta_1^g + \beta_2^g(sex_j) + \beta_3^g(smoke_j) + \beta_4^g(age_j) + \mu_j^g(subject) + \mu_k^g(hybridization) + \mu_l^g(label) + \varepsilon_{ijklm}^g \quad (2)$$

We identified differentially expressed probes as those with a statistically significant log-fold change (based on likelihood ratio tests). We computed p-values adjusted for multiple testing by controlling the false discovery rate (FDR) with the Benjamini-Hochberg procedure [25], using the *multtest* package in R. These FDR-adjusted p-values ≤ 0.05, the traditional experiment-wise type I error rate, were considered significant.

Pathway Enrichment Analysis

We used a method known as "structurally enhanced pathway enrichment analysis" (SEPEA_NT3) [26], which incorporates the

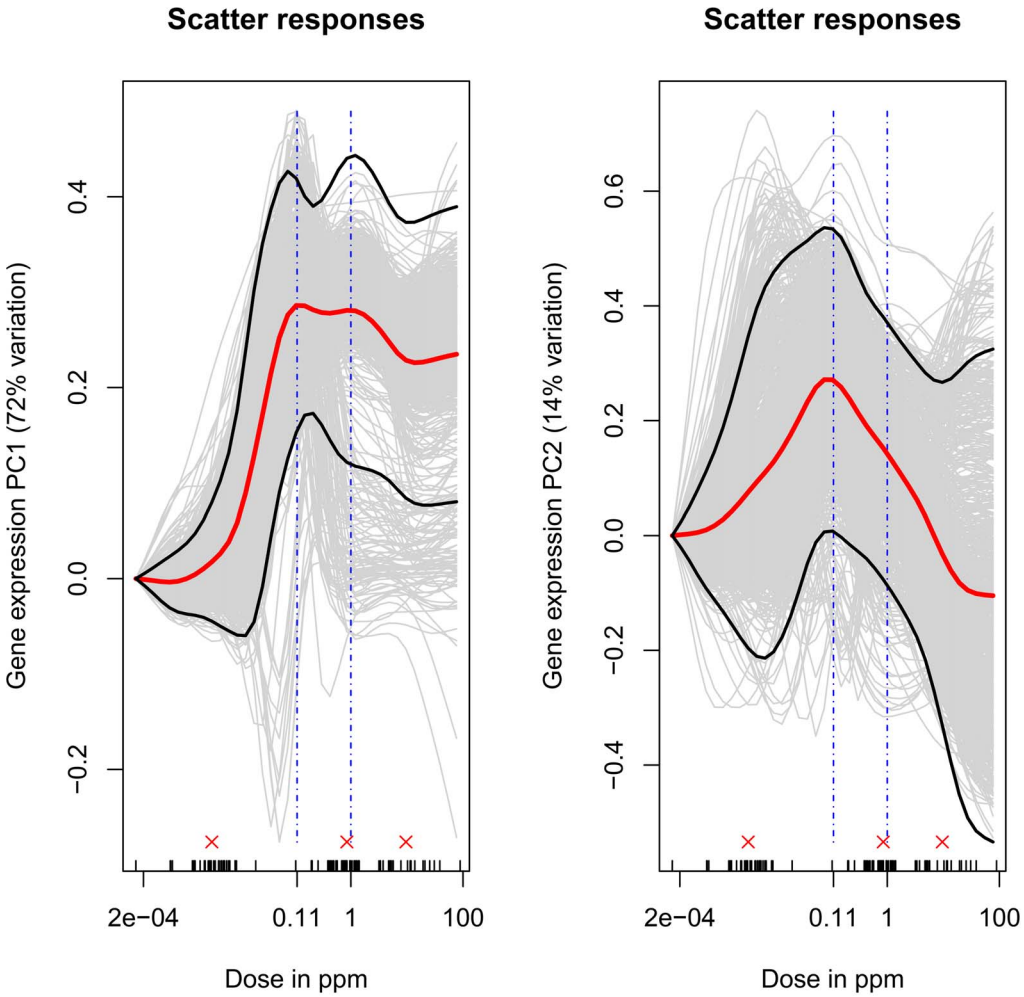

Figure 3. AML pathway-Principal Components-based response. The continuous fits in the two subplots use the first and second eigenvectors (that are slightly modified, see Material and Methods section) respectively from the eigenvector matrix, $Q^{p}_{s(n)}$ given in Equation (5). The elements of the individual eigenvectors are treated as the pathway response at the corresponding dose. The subscript *p* corresponds to the pathway under consideration and superscript *s* to a given bootstrap sample. The bold red line correspond the mean parameter estimates across the bootstrap samples and the bold black lines represent the corresponding 95% confidence intervals for the mean parameter estimates. The small vertical ticks on the x-axis denote doses to which one or more subjects in the study were exposed and consequently the doses for which data for all covariates under consideration were available. The three red 'x's above these ticks indicate the doses that there used to compare the rate of change of the marginal effect of benzene exposure from 0.001 to 1 ppm air benzene to the corresponding rate from 1 to 10 ppm air benzene.
doi:10.1371/journal.pone.0091828.g003

associated network information of KEGG (Kyoto Encyclopedia of Genes and Genomes) human biochemical pathways [19,27,28]. Unlike traditional pathway enrichment methods that treat pathways as sets of genes, SEPEA treats pathways as networks of interacting proteins and/or enzymes. The genes corresponding to the proteins in the signaling network are given more weight according to whether they are at the receptor or the terminating end of the pathway that typically signals for transcription in a number of genes. Further, pathways where the perturbed genes are close relative to each other on the associated network are modeled as being more likely to be affected than pathways where he perturbed genes occur further apart over the network. The significance obtained by SEPEA_NT3 was based on 10000 randomizations.

Bootstrap-SuperLearner

The SuperLearner [14] method is a theoretically optimal (relative to the so-called Oracle estimator) approach to model selection in a data adaptive manner. This method requires a set of different statistical learning algorithms that a user could consider as being appropriate models of the data. SuperLearner then uses a cross-validation-based loss function to estimate an optimal combination of predictions from the different input algorithms to produce model fits. The SuperLearner was used to fit $E\left[Y/_{A=a,W=w}\right]$ where Y, A and W have the same meaning as in the previous sections. Note in these analyses cell counts are included as additional confounders. The fits were computed in the SuperLearner package [29] implemented in the R statistical environment [23] with a choice of a 10-fold cross-validation-based loss function. The statistical learning algorithms used were random forests [30], multivariate adaptive regression splines [31],, bagging [32], Bayesian Generalized Linear Models [33], cforests [34,35,36], neural networks [37], loess regression [38] and support vector machines [39]. Different parameter settings for each of these algorithms in their respective R packages were used (see

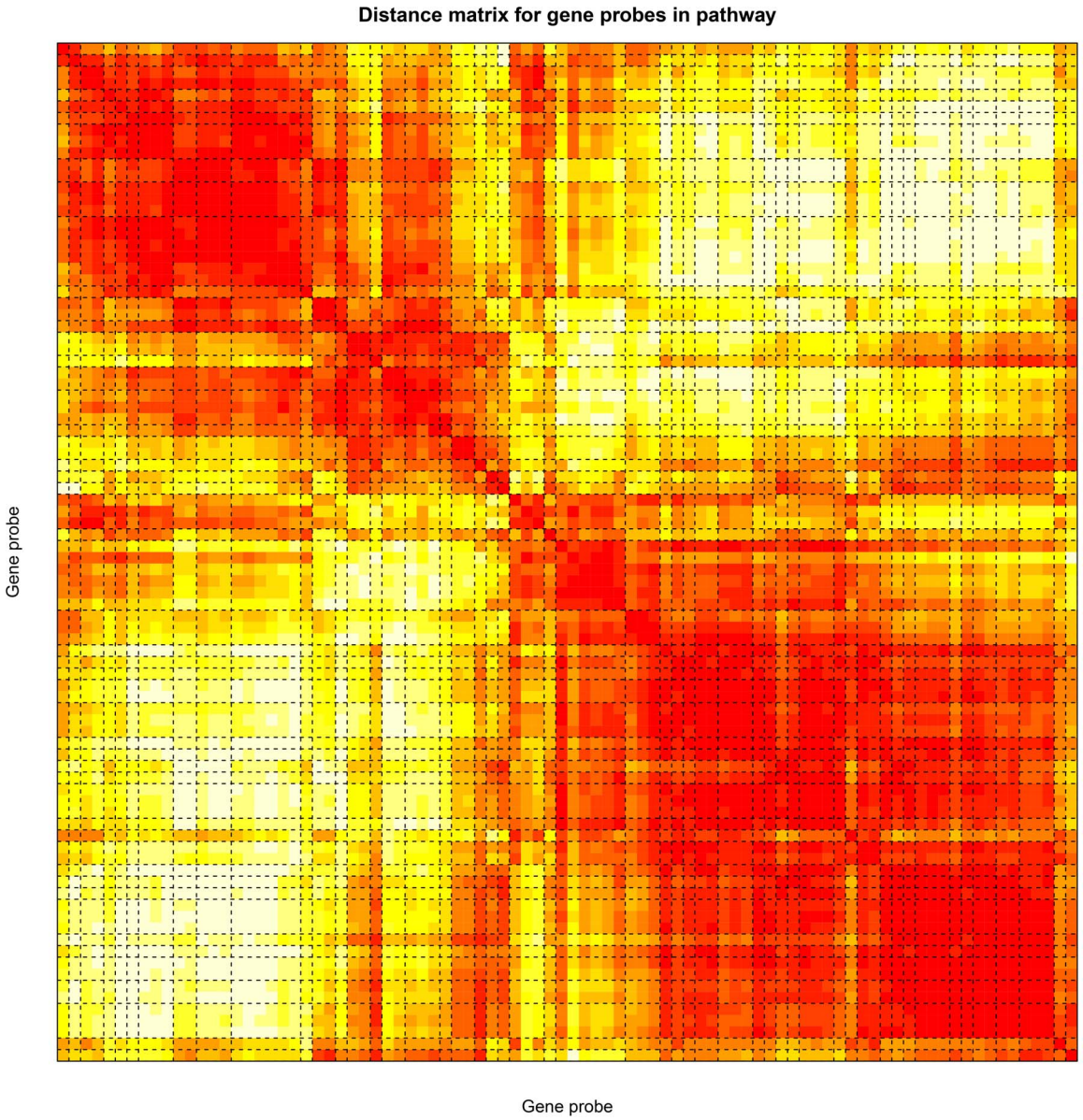

Figure 4. AML pathway-Clusters of probes/genes. Hierarchical cluster of the probes in the AML pathway. The probes are clustered based on the distance between the corresponding rows of the matrix, $X^p_\bullet$ given in Equation (6). The figure is a visual representation of the distance matrix between all the probes/genes in the pathway. The color of the $(i,j)^{th}$ position of the distance matrix is a measure of how close probes i and j are to each other based on their response across the dose range. The color ranges from white to red. The closer the pair of probes is two each other, the greater the intensity of red at the corresponding position. The dashed black lines correspond to boundaries of clusters of probes as determined by the HOPACH algorithm [47].
doi:10.1371/journal.pone.0091828.g004

Table S1)In total there were 32 learning algorithms. The reader interested in implementing the SuperLearner is referred to a vignette (http://cran.r-project.org/web/packages/SuperLearner/vignettes/SuperLearnerPresent.pdf) describing its implementation in R.

In order to determine the variability of the SuperLearner mean response estimates, a bootstrapping procedure was implemented. Let n denote the number of subjects and N_{BS} denote the number of bootstrap samples. 1000 bootstrap samples were chosen in all cases. Then for each bootstrap sample, n subjects are drawn randomly with replacement and the mean response is then estimated using the SuperLearner for each of the bootstrap samples.

The marginal association of a given response (expression of gene g) with the i^{th} dose of benzene exposure (corresponding to $A=a$) was then estimated by,

$$\begin{aligned}\Psi^g_{s(n),i} &= \Psi^g_{s(n)}(a) = \Psi^g\left(P_{s(n)}\right)(a)\\ &= P_{s(n)}\left(Q^g_{s(n)}(a,W) - Q^g_{s(n)}(0,W)\right)\\ &= \frac{1}{n}\sum_{j\in s(n)} Q^g_{s(n)}(a,W_j) - Q^g_{s(n)}(0,W_j)\end{aligned} \tag{3}$$

Where $s(n)$ represents a bootstrap sample, $P_{s(n)}$ represents the empirical distribution based on that sample, $Q^g_{s(n)}(a,W)$ the

Table 1. Median and 95% confidence interval (CI) estimates of the rate of change of marginal effect of benzene exposure below 1 ppm ($B^{i,1}/\beta^{g,1}$ – see equations (9) and (12)) and above 1 ppm ($B^{i,2}/\beta^{g,2}$ - see equations (10) and (13)) and the change in absolute rate of change of the marginal effects from below 1 ppm to above 1 ppm (D^i/δ^g – see equations (11) and (14)) for the first two principal components of the Acute Myeloid Leukemia pathway and six chosen genes of interest.

	$B^{i,1}/\beta^{g,1}$		$B^{i,2}/\beta^{g,2}$		Δ^i/δ^g	
Pathway/Gene	**Median**	**95% CI**	**Median**	**95% CI**	**Median**	**95% CI**
Acute Myeloid Leukemia: Principal Component 1	0.333	(0.008, 0.394)	−0.006	(−0.016, 0.002)	0.328	(0.019, 0.380)
Acute Myeloid Leukemia: Principal Component 2	0.106	(−0.244,0.349)	−0.024	(−0.040,0.010)	0.108	(−0.012,0.353)
RUNX1	0.07	(0.018, 0.158)	0.004	(−0.008, 0.025)	0.063	(0.010, 0.151)
FLT3	−0.079	(−0.204, −0.012)	0	(−0.005, 0.004)	0.078	(0.011, 0.202)
CEBPA	−0.877	(−1.11, −0.584)	0.03	(0.006, 0.069)	0.847	(0.561, 1.068)
LEF1	0.032	(−0.035, 0.120)	−0.009	(−0.019, −0.002)	0.025	(−0.008, 0.106)
CYP2E1	0.051	(−0.004, 0.147)	−0.002	(−0.007, 0.002)	0.049	(0.002, 0.146)
CYP2F1	0.002	(−0.033, 0.03)	−0.001	(−0.004, 0.001)	0.008	(−0.001, 0.038)

doi:10.1371/journal.pone.0091828.t001

estimate of $E\left[Y^g/A=a,W=w\right]$ based on the SuperLearner applied to $s(n)$. $A=0$ represents doses $\leq$0.11 ppm. So for every $A=a$(at which we calculated $\Psi^g_{s(n)}(a)$, which were at points separated 0.5 units apart on the $\log_2$ dose range), we get the average difference of the predicted $\log_2$ gene expression value (averaged across the W) and the predicted $\log_2$ gene expression value if the sample represented very low to no exposure.

All the subjects with undetectable benzene exposure levels in this study had predicted air benzene exposures less than 0.11 ppm. Under certain assumptions [40,41], this parameter of interest in Equation (3) also has also a simple causal interpretation, i.e., it represents the mean log fold change of a given gene's expression due to exposure to a given dose relative to those with exposure less than 0.11 ppm air benzene levels.

Biochemical Pathway Response

The derivation of the non-parametric estimate of the mean response of a biochemical pathway with an outcome of interest in the presence of confounders to its constituent gene expressions is a statistical problem that does not appear to have been addressed before in the literature. Examples of model-based approaches include those in [42,43] who proposed generalized linear model-based approaches to test for pathway association with a binary clinical outcome and survival times. We propose two ideas to provide summary responses of the expression of all the genes in a biochemical pathway, both of which use non-parametric estimates from the previously described Bootstrap-SuperLearner approach. The first idea is based on using principal component analysis (PCA) on the estimates from the SuperLearner of the expressions of genes in the pathway. Principal component analysis has been used in the past to model the pathway response [44,45]. The change in expressions of the genes due to benzene exposure is potentially confounded by other covariates in the study. Hence it will not be correct to perform a direct analysis of the expressions of the genes in the pathway in order to get a summary response. Therefore, PCA is performed on the SuperLearner-based non-parametric estimates of changes in gene expressions due to benzene exposure. The second idea utilizes a clustering analysis of these SuperLearner estimates in order to identify clusters of gene expression responses and medoid genes or particular genes that have responses that are representative of responses in the identified clusters. Clustering analysis of gene expression data [46] has been more or less standard for more than ten years.

Assume that the human biochemical pathways are identified by indices in the set $\{1,2,\ldots p,\ldots,N_{path}\}$ Let N_p denote the number of probes corresponding to genes involved in the given pathway identified by the index p. N_d is the number of points on the dose range of the exposed individuals where the SuperLearner [14] estimates are computed. Equally spaced points were chosen, 0.5 units apart on the logarithmic dose range.

In the first analysis, the dose-dependent responses were estimated by the first and second principal components (or first and second columns of the $Q^p_{s(n)}$ eigenvector matrix) of the covariance matrix, $\mathrm{cov}\left(\bar{X}^p_{s(n)}\right)$ created from the $N_p \times N_d$ matrix, $X^p_{s(n)}$ obtained for the SuperLearner [14] estimate of the parameters of interest (see Equations (4)-(6), where $\Lambda^p_{s(n)}$ represents diagonal matrix with the corresponding eigenvalues ordered in a decreasing manner). This was done for each bootstrap sample, s. Note, $\Psi^{\bullet}_{s(n),i}$ denotes the mean parameter of interest across all the probes at the i$^{\text{th}}$ dose level.

$$X^p_{s(n)} = \begin{pmatrix} \Psi^1_{s(n),1} & \cdots & \Psi^1_{s(n),N_d} \\ \vdots & \ddots & \vdots \\ \Psi^{Np}_{s(n),1} & \cdots & \Psi^{Np}_{s(n),N_d} \end{pmatrix} \quad (4)$$

$$\bar{X}^p_{s(n)} = \begin{pmatrix} \Psi^1_{s(n),1} - \Psi^{\bullet}_{s(n),1} & \cdots & \Psi^1_{s(n),N_d} - \Psi^{\bullet}_{s(n),N_d} \\ \vdots & \ddots & \vdots \\ \Psi^{Np}_{s(n),1} - \Psi^{\bullet}_{s(n),1} & \cdots & \Psi^{Np}_{s(n),N_d} - \Psi^{\bullet}_{s(n),N_d} \end{pmatrix} \quad (5)$$

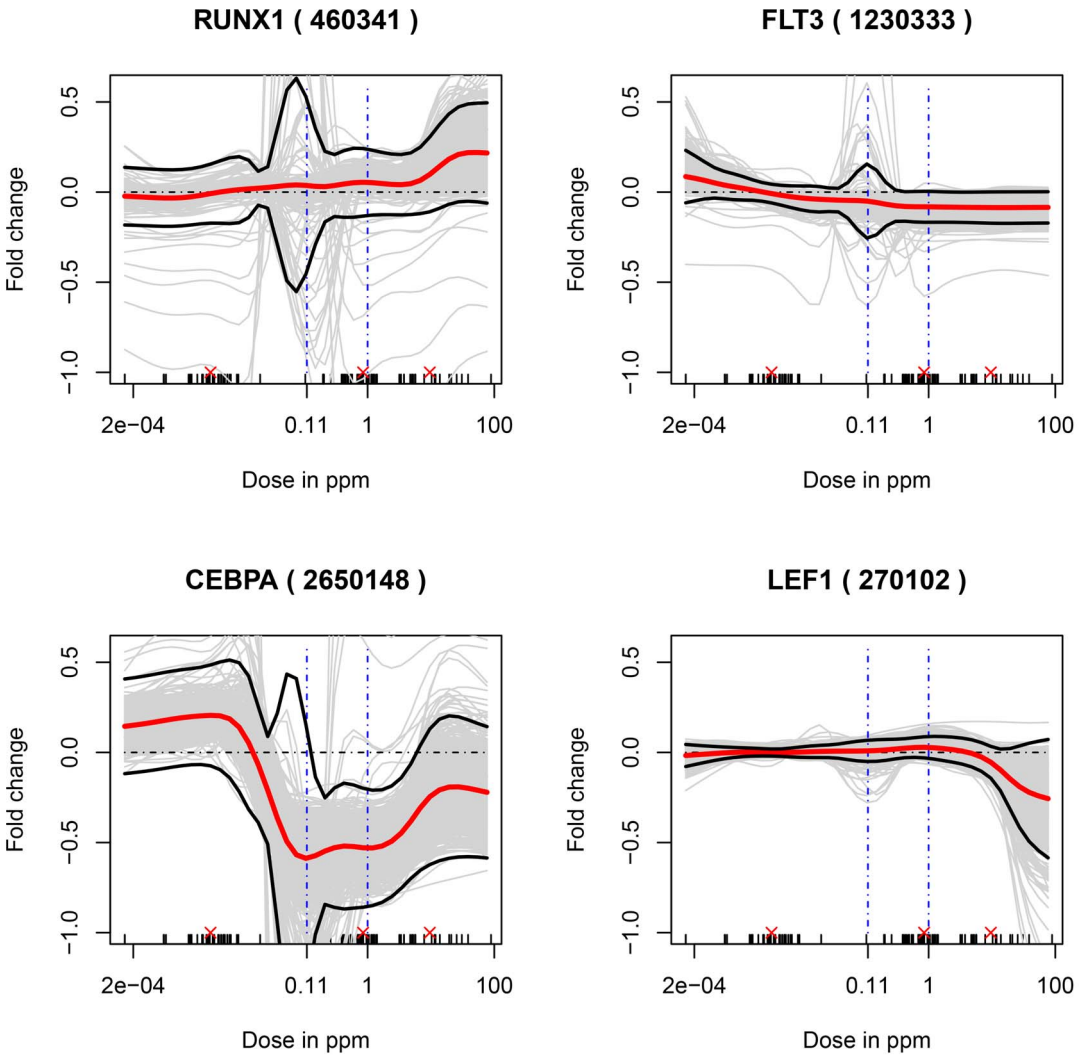

Figure 5. Responses of selected genes associated with the leukemia disease process. Non-parametric model fits to the expression response of the probes corresponding to six genes known to be associated with AML, with air-benzene concentrations in parts per million. Note the responses here are log fold-changes in expression. The dot-dashed horizontal line at a log fold change value equal to zero indicates the no-effect response. The gene names along with the corresponding probe id number on the microarray in parentheses are provided for each gene. The small vertical ticks on the x-axis denote doses to which one or more subjects in the study were exposed and consequently the doses for which data for all covariates under consideration were available. The three red 'x's above these ticks indicate the doses that there used to compare the rate of change of the marginal effect of benzene exposure from 0.001 to 1 ppm air benzene to the corresponding rate from 1 to 10 ppm air benzene.
doi:10.1371/journal.pone.0091828.g005

$$\mathrm{cov}\left(\bar{X}^{p}_{s(n)}\right) = \bar{X}^{p}_{s(n)} T \bar{X}^{p}_{s(n)} = Q^{p}_{s(n)} T \Lambda^{p}_{s(n)} Q^{p}_{s(n)} \quad (6)$$

Where

$$Q^{p}_{s(n)} = \left[q^{1,p}_{s(n)} \cdots q^{N_p,p}_{s(n)} \right] \quad (7)$$

And $q^{1,p}_{s(n)} \cdots q^{N_p,p}_{s(n)}$ are the N_p eigenvectors corresponding to the eigenvalues in the matrix $\Lambda^{p}_{s(n)}$. Each of the N_d elements of the first and second eigenvector was taken to represent the pathway response at the corresponding dose. In order to make comparisons of these eigenvectors across all bootstrap samples, two modifications were made to these eigenvectors. First, the first element of the (first or second) eigenvector for each of the bootstrap samples was normalized to zero. Second, if the sign of this normalized eigenvector responses at 0.1 ppm was negative, and then the negative of the elements of normalized eigenvector was plotted. This can be done because any scalar multiple of the reported eigenvector is an equally valid eigenvector for the given eigenvalue. We denote these modified eigenvectors by $\bar{q}^{1,p}_{s(n)}, \bar{q}^{2,p}_{s(n)}$ where for $i = 1,2$,

=

$$\bar{q}^{1,p}_{s(n)} = \left[\Phi^{i,p}_{s(n)}(a_1) \quad \cdots \quad \Phi^{i,p}_{s(n)}(a_{N_d}) \right] \quad (8)$$

The p^{th} pathway response at dose a_j is given by $\Phi^{i,p}_{s(n)}(a_j)$. Note by definition $\Phi^{i,p}_{s(n)}(a_1) = 0$.

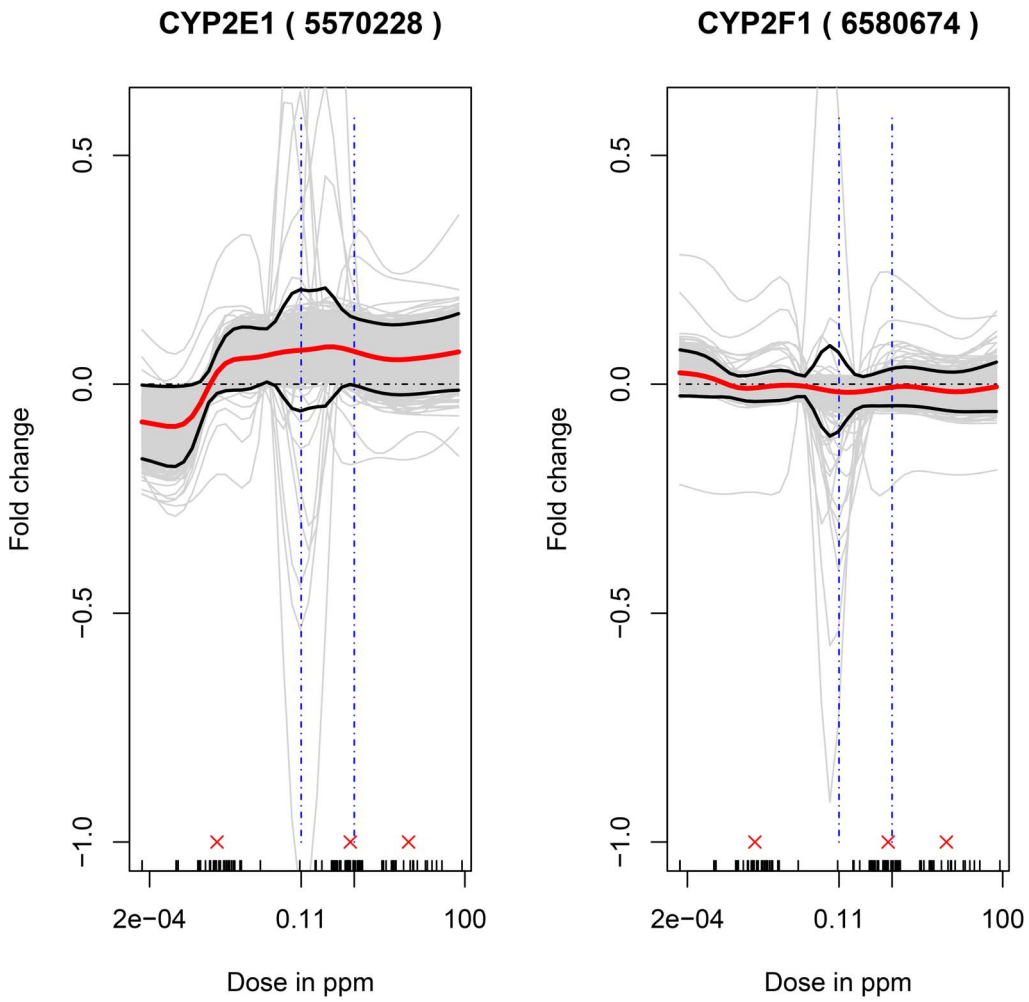

Figure 6. Response of two Cytochrome p450 genes associated with benzene metabolism. Non-parametric model fits to the expression response of the probes corresponding to two genes known to be associated with the metabolism of benzene, CYP2E1 and CYP2F1, with air-benzene concentrations in parts per million. Note the responses here are log fold-changes in expression. The dot-dashed horizontal line at a log fold change value equal of zero indicates the no-effect response. The gene names along with the corresponding probe id number on the microarray in parentheses are provided for each gene. The small vertical ticks on the x-axis denote doses to which one or more subjects in the study were exposed and consequently the doses for which data for all covariates under consideration were available. The three red 'x's above these ticks indicate the doses that there used to compare the rate of change of the marginal effect of benzene exposure from 0.001 to 1 ppm air benzene to the corresponding rate from 1 to 10 ppm air benzene.
doi:10.1371/journal.pone.0091828.g006

For each bootstrap sample, these modified normalized eigen-vector-based pathway response with dose is plotted as a smoothed cubic spline using a smoothing parameter of 0.5. The plots were made in the R statistical environment [23].

In the second analysis, the dose-dependent responses were presented as a clustered $(N_p \times N_p)$ distance matrix between the probes associated with the genes involved in the pathway. The clustering was performed using the HOPACH algorithm [47] in the package *hopach* [48] in the R statistical environment [23] where the distance between the two $N_d \times 1$ vectors, $\underline{\Psi^i_\bullet}$ and $\underline{\Psi^j_\bullet}$ associated with a pair of probes, i and j is measured by the cosine distance metric. Using this distance metric, HOPACH builds a hierarchical cluster of trees by recursively partioning the data set. Note the analyses here were performed on the matrix $X^p_\bullet$ that represents the average of $X^p_{s(n)}$ over all the bootstrap samples.

$$X^p_\bullet = \begin{pmatrix} \Psi^1_{\bullet,1} & \cdots & \Psi^1_{\bullet,N_d} \\ \vdots & \ddots & \vdots \\ \Psi^{Np}_{\bullet,1} & \cdots & \Psi^{Np}_{\bullet,N_d} \end{pmatrix} \tag{9}$$

From the clustering analyses, we also presented the responses of the genes identified as medoids for the six largest clusters as identified by the HOPACH algorithm [47].

The advantage of the principal component analyses of the pathway response is a one-picture summary response of the variability of the estimates of mean response of the pattern among a significant proportion of the genes in the pathway. However, this picture does not inform whether the genes are being over or under expressed at different levels of exposure. The plots of the medoid genes provide the sign of response of these chosen genes.

Estimates of the Change in the Rate of Change of Response

Characteristics of the shapes of the dose-response curves were obtained in terms of the change in the absolute rate of change of marginal effect of benzene exposure from below 1 ppm to above this level. For the two pathway responses ($i = 1,2$), the rate of change of the marginal effect response below 1 ppm for the bootstrap sample $s(n)$ is estimated by,

$$B^{i,1}_{s(n)} = \frac{\Phi^{i,p}_{s(n)}(a_2) - \Phi^{i,p}_{s(n)}(a_1)}{a_2 - a_1} \quad (10)$$

Where $a_2 = 1ppm$ and $a_1 = 0.001ppm$, $\Phi^{i,p}_{s(n)}(a_j)$ is as given in Equation (8). The rate of change of response above 1 ppm by,

$$B^{i,2}_{s(n)} = \frac{\Phi^{i,p}_{s(n)}(a_3) - \Phi^{i,p}_{s(n)}(a_2)}{a_3 - a_2} \quad (11)$$

Where $a_3 = 10ppm$ and the estimate of the change in the absolute rate of change of response, $D^{i}_{s(n)}$ is chosen to be

$$D^{i}_{s(n)} = |B^{i,1}_{s(n)}| - |B^{i,2}_{s(n)}| \quad (12)$$

Analogously the estimates of the change in the absolute rate of change of gene, g level responses, $\delta^{g}_{s(n)}$ are given by,

$$\beta^{i,1}_{s(n)} = \frac{\Psi^{i,p}_{s(n)}(a_2) - \Psi^{i,p}_{s(n)}(a_1)}{a_2 - a_1} \quad (13)$$

$$\beta^{i,2}_{s(n)} = \frac{\Psi^{i,p}_{s(n)}(a_3) - \Psi^{i,p}_{s(n)}(a_2)}{a_3 - a_2} \quad (14)$$

$$\delta^{i}_{s(n)} = |\beta^{i,1}_{s(n)}| - |\beta^{i,2}_{s(n)}| \quad (15)$$

Where $\Psi^{i,p}_{s(n)}(a_j)$ is given by Equation (3). The bootstrap samples of $D^{i}_{s(n)}$ and $\delta^{g}_{s(n)}$ are used to estimate the 95% confidence intervals for D^{i} and δ^{g}. Significant positive values of these estimates suggest supralinearity of the marginal effects of benzene exposure while significant negative values suggests sublinearity of these effects between 0.001 ppm and 10 ppm benzene levels.

Results

Predicted Air Benzene Exposure Levels in Non-occupationally Exposed Subjects

The air benzene exposure levels for the 42 control subjects [13] were predicted from their urinary unmetabolized benzene levels [17]. For 8 of the control subjects, exposure predictions were unavailable and these subjects were excluded from the non-parametric analyses. The predicted benzene exposure levels ranged from 1.4×10^{-4} ppm to 0.11 ppm, with 32 subjects predicted to have levels below 0.009 ppm.

Peripheral Blood Cell Counts as Potential Confounders of Gene Expression

Two linear mixed models of gene expression as a function of air benzene exposure were fitted to the data from 125 subjects who had been exposed to benzene in four previously chosen concentration ranges, $<<1$ ppm, <1 ppm, >5 ppm and $<$ 10 ppm, and ≥ 10 ppm,or were controls. One of the models was identical to that recently reported by us [13] and included the gender, age and smoking status of the subjects as potential confounders of gene expression. The second model was the same but also included the measured counts of different cell types present in the PBMC as additional confounders (see Equation (1)). The estimates of the fixed effects of this model are given in Table S2. The distribution of the estimates of the intra-class coefficients for each of the random effects do not change when moves from a model that does not include the PBMCs to one that does (see Figure S1). As shown in the Venn diagram in Figureô 2, there was a significant overlap (2183 genes) between the sets of genes identified as differentially expressed (FDR-adjusted p-value <0.05) by each linear model. Fisher's exact test estimated a p-value $< 2.2\times10^{-16}$ for the null hypothesis stating the independence between genes declared differentially expressed by the two linear models. When cell counts were incorporated as potential confounders, 821 out of 3004 genes identified as differentially expressed in the original model were no longer significant, while an additional 388 genes were found to be significant. There were also no major differences in the pathway enrichment values for all the KEGG human pathways using results from either model (Table S3). Pathway enrichment analyses were also done on the sets of genes that were commonly identified as differentially expressed genes (2183 genes) and uniquely by either of the models (821 or 388 genes) (see Table S3). The pearson correlation between the $\log_{10}$ transformed p-values across pathways for the set of genes uniquely identified by the model that did not incorporate cell count with the corresponding p-values for the set of genes commonly identified by both models is 0.06 (p-value = 0.32). The pearson correlation for the pathway p-values using the set of genes uniquely identified by the model that incorporated cell count with corresponding p-values for commonly identified set of genes is 0.22 (p-value = 0.0005).

Biochemical Pathway-based Responses

A dose-specific parameter of interest (defined in Equation (3)) was estimated non-parametrically for the expression of each probe/gene in the KEGG AML, B-cell receptor signaling, Toll-like receptor signaling, Steroid hormone biosynthesis, and Maturity onset of diabetes pathways, at points equally spaced 0.5 units apart on the logarithmic dose range. For a given gene at a chosen dose, this parameter is the expected log fold-change in the expression of the gene at that dose relative to the mean expression of subjects exposed to levels below 0.11 ppm. The parameters were estimated using the SuperLearner [14] that used 32 learning algorithms and the sampling distribution of these parameters was estimated via a bootstrapping procedure in which the parameters are re-estimated using random selection (with replacement) of the 125 subjects. The bootstrapping procedure was repeated 1000 times.

The first and second principal components of the estimated parameters for all genes in the AML pathway, evaluated across the entire study dose range, are shown in Figureô 3. Together, these two principal components captured 86% of the dose-dependent variation in expression of genes in the AML pathway. The first and second principal components of the dose-response parameters for all genes in the B-cell receptor signaling, Toll-like receptor signaling, Steroid hormone biosynthesis and Maturity onset of diabetes pathways are shown in figure S2, respectively. Visually, the mean estimates of the first principal components of the responses look similar across the five chosen pathways and suggest

supra-linear responses. This is quantitatively reinforced by the fact that the estimates of the change in the absolute rate of change of marginal effect of benzene exposure from below 1 ppm to above 1 ppm for the first principal component of each of the five pathways (see Equation (12)) are significantly positive (p-value $<$ 0.05) (see Table 1 and Table S4). There are suggestions of responses at doses below 0.1 ppm, and exposures of around 0.1 ppm and 1 ppm appear to be inflection points for at least three (AML, Toll-like receptor signaling and Maturity Onset of Diabetes) of the five pathways. The mean estimates of the second principal components are also similar across the five pathways, with exposures around 0.1 ppm appearing to represent inflection points for these responses. The responses of the probes in the pathway were clustered using HOPACH [49] (Figureô 4 and figure S3), and the results confirm that many sets of genes exhibit similar response patterns.

Expression-based Responses of Chosen Genes of Interest

A small set of genes was chosen based either on their known association with leukemogenesis or because they code for enzymes putatively associated with the metabolism of benzene to its toxic metabolites. In the first class, oncogenes (RUNX1, FLT3, LEF1) or tumor suppressors (RUNX1, CEBPA) implicated in leukemia [50,51,52,53,54,55] were selected. RUNX1 has been identified as both an oncogene and a tumor suppressor gene [56,57]. The responses (in terms of log-fold changes) of these genes are plotted in Figureô 5. Among these genes, RUNX1, LEF1 and CEBPA were differentially expressed (FDR$<$0.05) based on the linear model in Equation (1) across four binned dose ranges. Based on the position of the line corresponding to no change (i.e., log fold change equals zero) relative to the 95% confidence interval of the estimated fold change at a chosen dose, the expressions of FLT3 and CEBPAare all down regulated on exposure to levels of air benzene above around 0.1 ppm. RUNX1, FLT3 and CEBPA display profiles that are supralinear as captured by the significant positive values of the change in the absolute rate of change of marginal effect of benzene exposure from below 1 ppm to above 1 ppm (see Equation (15) and Tableô 1). The responses of the expression of two genes, CYP2E1 and CYP2F1 (enzymes putatively associated with the metabolism of benzene to its toxic metabolites), are shown in Figureô 6. CYP2E1 is known to be involved with the metabolism of benzene [58,59,60] and CYP2F1 has been hypothesized to be associated with benzene metabolism [3]. CYP2E1 expression was increased at levels of air benzene concentration as low as 0.01 ppm, while expression of CYP2F1 was largely unaltered.CYP2E1, but not CYP2F1, was found to be differentially expressed (FDR$<$0.05) based on the linear model in Equation (1) across four binned dose ranges. The dose-response of CYP2E1 is supralinear (see Equation (15) and Tableô 1).

Discussion

The analyses in this study sought to further characterize the dose-dependency of changes in gene expression associated with occupational exposure to benzene that we reported recently [13] and to extend them into the ambient environmental range by estimating exposures for the non-occupationally exposed 'controls'. We further extended the analyses to include PBMC subset cell counts as potential confounders of expression.

Significant overlap was seen for the majority of genes identified as differentially expressed using the parametric linear models, regardless of whether or not PBMC cell counts were incorporated as potential confounders. Genes that did not remain differentially expressed after incorporation of cell counts as confounders could be indirectly related to benzene-induced cell count decrements. One plausible example is the CD44 gene; it encodes a marker for CD4 and CD8 cells, both of which were reduced in number in benzene-exposed individuals [7]. However, the estimation of additional parameters in the model to incorporate cell counts as confounders resulted in loss of statistical power to identify genes as differentially expressed. Conversely, the incorporation of cell counts in the model may improve the model fit and subsequently increase the ability of the model to detect true dose-specific changes in gene expression. This is partly suggested by the significant but relatively small correlation (0.22) of the pathway p-values using the set of genes uniquely identified by the model incorporating cell counts with the pathway p-values using the set of genes commonly identified by both models. In either case, the fact that a significant overlap was observed implies that the majority of the changes in gene expression are not directly mediated through the hematotoxicity of benzene.

We defined the dose-dependent effects on gene expressions as our parameters of interest as a marginal effect of benzene exposure. The definition of the parameter in dose-response studies of this kind is to the best of our knowledge novel in the toxicology literature – that is, as the marginally adjusted curve of the mean outcome versus exposure. This is extremely important in itself, because when measuring how dose-response affects a population, one should estimate a population level dose response parameter. However, what is typically done is to estimate the dose response in a parametric or semi-parametric model, where the resulting curve represents potentially a different curve for every covariate group. To avoid this problem, the predominant approach is to simply make the simplifying, but erroneous assumption, that all groups have the same dose response curve). We wanted to make no such bias inducing assumptions, and so we estimated these dose response curves in a nonparametric model, using optimal data-adaptive methods (SuperLeanrer [14]) for estimating the relationship of expression to exposure and the confounders. Thus, we have approached a problem typically approached using ad hoc methods, based on arbitrary statistical modeling assumptions, and used methods based on what is truly known about the relationship of expression to the covariates including exposure, that is nearly nothing, and derived optimal (the SuperLearner is not ad hoc, but based on the theorem of the Oracle Inequality [14]) estimators respecting the underlying knowledge. We use a bootstrapping approach in order to estimate the sampling variability of these estimates. Note that these estimates are essentially derived from data-adaptive methods in a very large (nearly non-parametric) model, where the number of assumptions made about the probability generating distributions is minimal, particularly as compared to standard approaches using parametric models.

The five pathways chosen for analyses were selected based on the results to our earlier analyses [13] with the same gene expression data on the same set of KEGG [19,27,28] human pathways. Ideally, we should have performed these analyses while being agnostic to the potential biochemical pathways being targeted. This would mean analyzing all 22177 probes on the microarray by the proposed non-parametric methodology. However, we chose to analyze the 5 pathways in part because the proposed non-parametric methodology requires significant computing power- one bootstrap sample run of the dose response of a given probe/gene using the SuperLearner [14] that ran 32 learning algorithms took around 20 seconds to run on a 4-core linux machine with around 1 GHz cpu and 16 GB RAM. We should note that the associations of benzene exposure are being made with biochemical pathways of diseases (AML and Maturity

Onset of Diabetes) and not with the diseases directly. These biochemical pathways represent a summary of the literature on the specific disease pathogenesis. Further the exact definition of a specific biochemical pathway in terms of its constituents and associated interactions will be consistent though not the same across different pathway databases. Therefore our choices of the pathways were from the same set of pathways analyzed before albeit with the same data – our goal here being a better characterization of the dose responses over the entire continuous range of exposures. The AML pathway was of particular interest because of the established association of benzene exposure with leukemia incidence [1,2]. The other two pathways (B-cell receptor and Toll-like receptor signaling) were randomly chosen from the set of pathways which displayed significant dose response over the range of benzene exposures. Similarly the Steroid Hormone Biosynthesis and Maturity onset of diabetes pathways were chosen from the list of pathways that did not display statistical significant responses.

The mean estimates of first and second principal components of the dose-response relationships determined by the our method method for all genes in the five chosen pathways (AML, B-cell receptor signaling, Toll-like receptor signaling, Steroid hormone biosynthesis and Maturity onset of diabetes) showed apparent similarities and similar inflection points were observed for several of the pathways. Since two of the analyzed pathways, Steroid hormone biosynthesis and Maturity onset of diabetes, are presumably unrelated to benzene, the noted similarities in response implies that the changes in expression of genes in these pathways are real effects though they were not large enough to provide statistical significance for modulation at the pathway level. Consistency of the shapes of the responses across the five chosen pathways may be a consequence of the correlation among gene expression levels on a system-wide basis through coordinated transcriptional regulation. However, analyses of the binding sites in the promoter regions of the differentially expressed genes (across doses, as determined by the linear model in Equation (1)) did not reveal enrichment of any transcription factor binding sites (data not shown).

Together, these similarities support the plausibility of the observed supra-linear dose-responses. In addition supralinearity is quantitatively supported by positive values of the parameter that captures the change in the absolute rate of change of marginal effect of benzene exposure from below 1 ppm to above 1 ppm (see Equations (12) and (15), Tableô 1). This adds to the literature of observed supra-linear responses associated with benzene exposure – see for example the response of benzene oxide-albumin adduct formation with benzene exposure [61], the dose related production of benzene metabolites [17] and the relative risk of leukemia with benzene exposure [10].

Our statistical tests for supralinearity are based on comparing the rate of change of the marginal effect of exposure below 1 ppm (0.001–1 ppm) benzene to the rate of change of this marginal effect above 1 ppm (1–10 ppm). We don't perform statistical tests for supra-linearity of the overall dose response curve. Testing for supra-linearity is a very subtle issue, since any data adaptive approach, compared to some model in a goodness of fit test, will always win out asymptotically (any null model will be not perfectly right, so as sample size grows, and the data-adaptive approach will favor more highly parameterized models to create a better fit, any improvement over the null becomes statistically significant). Thus, any conclusion made from a test in this context is dubious, since the asymptotic p-value will always go to 0. Therefore, we used an approach based on confidence intervals of the overall dose-response curve, which do not have this particular pathology. Several genes of interest were chosen for examination based on their association with leukemogenesis or benzene metabolism. The observed significant decrease in CEBPA expression at benzene levels of around 0.1 ppm may be important in light of the fact that reduced CEBPA gene expression has been associated with increased risk of leukemia [55]. Changes in the expression of CYP2E1 were observed at levels of air benzene concentration as low as 0.01 ppm. As benzene metabolism occurs principally in the liver [62] and also in the lung [63,64], with secondary metabolism occurring in the bone marrow [65,66,67], the implication of altered expression of CYP2E1 in peripheral blood is not entirely clear.

In order to permit phenotypic anchoring of the observed dose-dependent changes in gene expression, the responses of counts of B-cells, white blood cells and the ratio of the counts of CD4 cells to CD8 cells, are estimated using the Bootstrap-SuperLearner approach (data not shown). The observed decreases in these cell counts have been previously reported [7]. Mean changes in gene expression can thus be associated with corresponding changes in mean cell counts. For example, a 0.75 expected fold change of CEBPA gene expression at 1 ppm of air benzene would be associated with a mean decrease of 600 white blood cells/µl or a decrease of 0.2 in the ratio of CD4 to CD8 cells.

In summary, this work presents a new approach, which is the combination of the choice of a statistical parameter to be estimated, the methods used to estimate the data generating distribution (our parameter is only a targeted part of that distribution), and rigorous and robust methods for deriving inference, applied to the scientific question of estimating the dose-dependent biological responses resulting from exposure to benzene in the air. This work extends our previous analyses of benzene-induced differential gene expression in occupationally exposed workers and demonstrates that the differential expression of the majority of genes is independent of changes in cell counts of various blood cell types; that many differentially expressed genes and disease-relevant pathways display an apparently supra-linear response; and, that benzene alters these pathways and genes at exposure levels as low as 0.1 ppm. However, limitations in the statistical models and in the interpretation of some of these findings suggest that studies with a larger number of samples from individuals exposed to benzene in the low-dose range between 0.01 and 1 ppm are needed to clarify and confirm our interpretations. More precise exposure measurement and/or estimation in the low-dose region are needed to clarify the nature of the dose-response relationship of gene alteration in the low-dose range.

Supporting Information

Figure S1 Distribution of intra-class coefficients of the chip, subject, labeling and hybridization random effects.
(PDF)

Figure S2 Pathways-Principal Components-based Response. Non-parametric model fits to the marginal association of the expression of the probes corresponding to the genes involved in the a. B-cell receptor signaling, b.Toll-like receptor signaling, c. Steroid Hormone bio-synthesis and d.Maturity onset of diabetes pathways with air-benzene concentrations in parts per million. The continuous fits in the two subplots use the first and second eigenvectors (that are slightly modified, see Material and Methods section) respectively from the eigenvector matrix, $Q^p_{s(n)}$ given in Equation (5). The elements of the individual eigenvectors are treated as the pathway response at the corresponding dose. The

subscript p corresponds to the pathway under consideration and superscript s to a given bootstrap sample. The bold red line correspond the mean parameter estimates across the bootstrap samples and the bold black lines represent the corresponding 95% confidence intervals for the mean parameter estimates. The small vertical ticks on the x-axis denote doses to which one or more subjects in the study were exposed and consequently the doses for which data for all covariates under consideration were available. The three red 'x's above these ticks indicate the doses that there used to compare the rate of change of the marginal effect of benzene exposure from 0.001 to 1 ppm air benzene to the corresponding rate from 1 to 10 ppm air benzene.
(PDF)

Figure S3 Pathways-Clusters of probes/genes. Non-parametric model fits to the marginal association of the expression of the probes corresponding to the genes involved in the a. B-cell receptor signaling, b.Toll-like receptor signaling, c. Steroid Hormone bio-synthesis and d.Maturity onset of diabetes pathways with air-benzene concentrations in parts per million. The probes are clustered based on the distance between the corresponding rows of the matrix, $X^p_\bullet$ given in Equation (6). The figure is a visual representation of the distance matrix between all the probes/genes in the pathway. The color of the $(i,j)^{th}$ position of the distance matrix is a measure of how close probes i and j are to each other based on their response across the dose range. The color ranges from white to red. The closer the pair of probes is two each other, the greater the intensity of red at the corresponding position. The dashed black lines correspond to boundaries of clusters of probes as determined by the HOPACH algorithm [47].
(PDF)

Table S1 List of supervised learning algorithms.
(XLSX)

Table S2 Fixed effects estimates for the mixed model in Equation (1).
(XLSX)

Table S3 p-Values for KEGG pathways. The p-values were computed using the *SEPEA_NT3* procedure [26] based on results of differential from expression (in at least one of the four benzene exposure groups) from the linear mixed models with (L1) and without (L0) using the blood cell counts as potential confounders of gene expression. Also listed are the p-values obtained the KEGG pathway enrichment using genes commonly identified by both models, unique to the model (L0) and unique to the model (L1).
(XLSX)

Table S4 Median and 95% confidence interval (CI) estimates of the rate of change of marginal effect of benzene exposure below 1 ppm ($B^{i,1}/\beta^{g,1}$– see equations (10) and (13)) and above 1 ppm ($B^{i,2}/\beta^{g,2}$ - see equations (11) and (14)) and the change in absolute rate of change of the marginal effects from below 1 ppm to above 1 ppm (D^i/δ^g – see equations (12) and (15)) for the first two principal components of the for the B-cell receptor signaling, Toll-like receptor signaling, Steroid hormone synthesis and Maturity onset of diabetes pathways.
(XLSX)

Acknowledgments

Disclaimer: The views expressed in this manuscript are those of the authors and do not necessarily represent opinion or policy of the US Environmental Protection Agency.

Author Contributions

Analyzed the data: RT AEH. Designed the study: MTS NR LZ QL. Developed the statistical methods: RT AEH. Performed the microarray experiments: CMM. Performed the exposure assessment: RV SMR. Prepared the manuscript draft: RT. Provided important intellectual and editorial input: CMM AEH LZ MTS AEH NR QL RV SMR JJ KZG BRS. All authors Approved the final manuscript: RT AEH CMM LZ SMR QL NR RV KZG JJ BRS MTS.

References

1. Khalade A, Jaakkola MS, Pukkala E, Jaakkola JJK (2010) Exposure to benzene at work and the risk of leukemia: a systematic review and meta-analysis. Environmental Health 9: 31.
2. Steinmaus C, Smith AH, Jones RM, Smith MT (2008) Meta-analysis of benzene exposure and non-Hodgkin lymphoma: biases could mask an important association. Occupational and environmental medicine 65: 371.
3. Rappaport SM, Kim S, Lan Q, Vermeulen R, Waidyanatha S, et al. (2009) Evidence that humans metabolize benzene via two pathways. Environmental health perspectives 117: 946.
4. Smith MT, Zhang L, McHale CM, Skibola CF, Rappaport SM (2011) Benzene, the Exposome and Future Investigations of Leukemia Etiology. Chemico-Biological Interactions.
5. Zhang L, McHale CM, Rothman N, Li G, Ji Z, et al. (2010) Systems biology of human benzene exposure. Chem Biol Interact 184: 86–93.
6. McHale CM, Zhang L, Smith MT (2012) Current understanding of the mechanism of benzene-induced leukemia in humans: implications for risk assessment. Carcinogenesis 33: 240–252.
7. Lan Q, Zhang L, Li G, Vermeulen R, Weinberg RS, et al. (2004) Hematotoxicity in workers exposed to low levels of benzene. Science 306: 1774.
8. Rappaport SM, Waidyanatha S, Yeowell-O'Connell K, Rothman N, Smith MT, et al. (2005) Protein adducts as biomarkers of human benzene metabolism. Chemico-Biological Interactions 153: 103–109.
9. Rappaport SM, Yeowell-O'Connell K, Smith MT, Dosemeci M, Hayes RB, et al. (2002) Non-linear production of benzene oxide-albumin adducts with human exposure to benzene. Journal of Chromatographyô B 778: 367–374.
10. Vlaanderen J, Portengen L, Rothman N, Lan Q, Kromhout H, et al. (2010) Flexible meta-regression to assess the shape of the benzene-leukemia exposure-response curve. Environ Health Perspect 118: 526–532.
11. Lan Q, Vermeulen R, Zhang L, Li G, Rosenberg PS, et al. (2006) Benzene Exposure and Hematotoxicity: Response. Science 312: 998–998.
12. Qu Q, Shore R, Li G, Jin X, Chen LC, et al. (2002) Hematological changes among Chinese workers with a broad range of benzene exposures. Amô Jô Ind Med 42: 275–285.
13. McHale CM, Zhang L, Lan Q, Vermeulen R, Li G, et al. (2011) Global Gene Expression Profiling of a Population Exposed to a Range of Benzene Levels. Environmental Health Perspectives 119: 628–640.
14. van Der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Statistical applications in genetics and molecular biology 6: 25.
15. Bolen CR, Uduman M, Kleinstein SH (2011) Cell subset prediction for blood genomic studies. BMC bioinformatics 12: 258.
16. Vermeulen R, Li G, Lan Q, Dosemeci M, Rappaport SM, et al. (2004) Detailed exposure assessment for a molecular epidemiology study of benzene in two shoe factories in China. Annals of Occupational Hygiene 48: 105.
17. Kim S, Vermeulen R, Waidyanatha S, Johnson BA, Lan Q, et al. (2006) Using urinary biomarkers to elucidate dose-related patterns of human benzene metabolism. Carcinogenesis 27: 772.
18. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic acids research 36: D480.
19. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27.
20. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic acids research 34: D354.
21. Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics: 963–974.
22. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307.
23. Team RDC (2004) R: a language and environment for statistical computing. Rô foundation for Statistical Computing.
24. Bates D, Maechler M, Dai B (2008) lme4: linear mixed-effects models using S4 classes. R package version 0.999375–33. R Foundation for Statistical Computing. Vienna, Austria. lme4 r-forge r-project org/i.
25. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Seriesô Bô (Methodological): 289–300.

26. Thomas R, Gohlke J, Stopper G, Parham F, Portier C (2009) Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. Genome Biology 10: R44.
27. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic acids research 36: D480.
28. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic acids research 34: D354.
29. Polley EC (2010) SuperLearner: Super Learner Prediction. Rô package version 11–18. Available: http://wwwstatberkeleyedu/~ecpolley/SL/.
30. Breiman L (2001) Random forests. Machine learning 45: 5–32.
31. Friedman JH (1991) Multivariate adaptive regression splines. The annals of statistics: 1–67.
32. Breiman L (1996) Bagging predictors. Machine learning 24: 123–140.
33. Gelman A, Su YS, Yajima M, Hill J, Pittau MG, et al. (2010) arm: Data analysis using regression and multilevel/hierarchical models. Rô package version: 1.3–02.
34. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ (2006) Survival ensembles. Biostatistics 7: 355.
35. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC bioinformatics 9: 307.
36. Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8: 25.
37. Haykin S (1999) Neural networks: a comprehensive foundation: Prentice hall.
38. Cleveland W, Grosse E, Shyu W, Chambers J, Hastie T (1991) Statistical models in S. Wadsworth and Brooks/Cole, Pacific Grove, Ch Local regression models: 309–376.
39. Hearst MA, Dumais S, Osman E, Platt J, Scholkopf B (1998) Support vector machines. Intelligent Systems and their Applications, IEEE 13: 18–28.
40. Gill RD, Robins JM (2001) Causal inference for complex longitudinal data: the continuous case. Annals of Statistics: 1785–1811.
41. van Der Laan MJ, Rose S (2011) Targeted Learning: Causal Inference for Observational and Experimental Data: Springer Verlag.
42. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, Van Houwelingen HC (2005) Testing association of a pathway with survival using gene expression data. Bioinformatics 21: 1950–1957.
43. Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
44. Chen X, Wang L, Smith JD, Zhang B (2008) Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics 24: 2474–2481.
45. Ma S, Kosorok MR (2009) Identification of differential gene pathways with principal component analysis. Bioinformatics 25: 882–889.
46. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863.
47. van der Laan MJ, Pollard KS (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. Journal of Statistical Planning and Inference 117: 275–303.
48. Pollard KS, Wall G, van der Laan MJ (2010) hopach: Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH). Rô package version 2100. Available: http://CRANR-projectorg/package = hopach.
49. van der Laan M, Pollard K (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. Journal of Statistical Planning and Inference 117: 275–303.
50. Choudhary C, Müller-Tidow C, Berdel WE, Serve H (2005) Signal transduction of oncogenic Flt3. International journal of hematology 82: 93–99.
51. Lorsbach RB, Downing JR (2001) The role of the AML1 transcription factor in leukemogenesis. International journal of hematology 74: 258–265.
52. Metzeler KH, Heilmeier B, Edmaier KE, Rawat VP, Dufour A, et al. (2012) High expression of lymphoid enhancer-binding factor-1 (LEF1) is a novel favorable prognostic factor in cytogenetically normal acute myeloid leukemia. Blood 120: 2118–2126.
53. Mizuki M, Schwäble J, Steur C, Choudhary C, Agrawal S, et al. (2003) Suppression of myeloid transcription factors and induction of STAT response genes by AML-specific Flt3 mutations. Blood 101: 3164–3173.
54. Steffen B, Muller-Tidow C, Schwable J, Berdel WE, Serve H (2005) The molecular pathogenesis of acute myeloid leukemia. Critical reviews in oncology/hematology 56: 195–221.
55. Van Doorn SBVW, Khosrovani CE, Meijer J, Van Oosterhoud S, Van Putten W, et al. (2003) Biallelic mutations in the CEBPA gene and low CEBPA expression levels as prognostic markers in intermediate-risk AML. Hematolô J 4: 31–40.
56. Silva FP, Morolli B, Storlazzi CT, Anelli L, Wessels H, et al. (2003) Identification of RUNX1/AML1 as a classical tumor suppressor gene. Oncogene 22: 538–547.
57. Wotton S, Stewart M, Blyth K, Vaillant F, Kilbey A, et al. (2002) Proviral insertion indicates a dominant oncogenic role for Runx1/AML-1 in T-cell lymphoma. Cancer research 62: 7181–7185.
58. Koop DR, Laethem CL, Schnier GG (1989) Identification of ethanol-inducible P450 isozyme 3a (P450IIE1) as a benzene and phenol hydroxylase. Toxicology and applied pharmacology 98: 278–288.
59. Nedelcheva V, Gut I, Souček P, Tichavska B, Týnkova L, et al. (1999) Metabolism of benzene in human liver microsomes: individual variations in relation to CYP2E1 expression. Archives of toxicology 73: 33–40.
60. Powley MW, Carlson GP (2000) Cytochromes P450 involved with benzene metabolism in hepatic and pulmonary microsomes. Journal of Biochemical and Molecular Toxicology 14: 303–309.
61. Rappaport SM, Yeowell-O'Connell K, Smith MT, Dosemeci M, Hayes RB, et al. (2002) Non-linear production of benzene oxide–albumin adducts with human exposure to benzene. Journal of Chromatographyô B 778: 367–374.
62. Sammett D, Lee EW, Kocsis JJ, Snyder R (1979) Partial hepatectomy reduces both metabolism and toxicity of benzene. Jô Toxicol Environ Health 5: 785–792.
63. Powley MW, Carlson GP (2002) Benzene metabolism by the isolated perfused lung. Inhal Toxicol 14: 569–584.
64. Sheets PL, Yost GS, Carlson GP (2004) Benzene metabolism in human lung cell lines BEAS-2B and A549 and cells overexpressing CYP2F1. Jô Biochem Mol Toxicol 18: 92–99.
65. Andrews LS, Sasame H, Gillette JR (1979) 3H-Benzene metabolism in rabbit bone marrow. Life sciences 25: 567–572.
66. Subrahmanyam VV, Doane-Setzer P, Steinmetz KL, Ross D, Smith MT (1990) Phenol-induced stimulation of hydroquinone bioactivation in mouse bone marrow in vivo: possible implications in benzene myelotoxicity. Toxicology 62: 107–116.
67. Subrahmanyam VV, Kolachana P, Smith MT (1991) Hydroxylation of phenol to hydroquinone catalyzed by a human myeloperoxidase-superoxide complex: possible implications in benzene-induced myelotoxicity. Free radical research communications 15: 285–296.